Metagenome Analysis With MG-RAST

Folker Meyer, PhD

Argonne National Laboratory
and
University of Chicago

http://metagenomics.anl.gov

Palm Springs, March 2013

Acknowledgements

Team:

- Dion Antonopoulos
- Daniela Bartels
- Jared Bischof
- Dan Braithewaite
- Sarah O'Brien
- Adina Chuang-Howe
- Narayan Desai
- Mark Domanus
- Mark d' Souza
- Katya Drybinski
- Elizabeth M. Glass
- Wolfgang Gerlach
- Kim M. Handley
- Travis Harrison
- Kevin Keegan
- Tobias Paczian
- Hunter Matthews
- Sarah Owens
- Wei Tang
- Will Trimble
- Andreas Wilke
- Jared Wilkening

Major Collaborators:

- A. Arkin (Berkeley)
- E. Chang (UChicago)
- Dawn Field (Oxford)
- F.-O. Glöckner (MPI Bremen)
- Jack Gilbert (Argonne)
- Jeff Grethe (CalIT2, CAMERA)
- Sarah Hunter / Guy Cochrane (EBI)
- Ken Kemner (Argonne)
- Rob Knight (Colorado)
- Nikos Kyrpides (DOE JGI)
- J. Tiedje (MSU)
- Owen White (UMaryland, HMP DACC)

Microbial community science

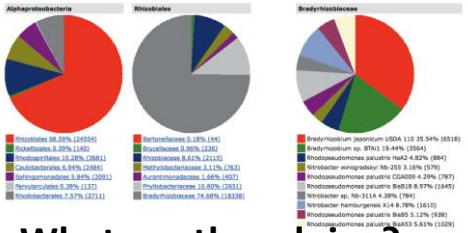
Who are they?

➔ **Discovery of novel functions**

Environmental clone libraries ("functional metagenomics")

- Sanger sequencing of BAC clones with env. DNA

→low throughput but supports in vitro screens



What are they doing?

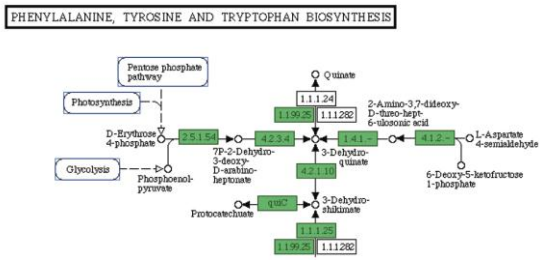
➔ **Ecology**

Amplicon studies (single gene studies, 16s rDNA)

- Sequencing of PCR amplified ribosomal genes

→ sequence quality limited (→ rare biosphere debate)

→often can't distinguish between individual strains



➔**Ecology and Discovery**

Shotgun metagenomics

- "random shotgun DNA sequencing applied directly to environmental samples"

Data → qc → genes → proteins → functional profiles → your science
→ organisms → taxonomic profiles

What is MG-RAST

- Web based
  - Upload, process, share and publish microbial community data
  - Upload for reads and assemblies
  - Extensive QC
- Automated processing and analysis of
  - ITS / 16s / 18s
  - shotgun metagenome
  - meta transcriptome
- Comparison
  - Functional
  - Taxonomic
  - Using precomputed profiles
- Subset retrieval
  - All reads for *E.coli*
  - All reads for Lysine Biosynthesis
  - All reads hitting dnaA

MG-RAST
metagenomics analysis server

# of metagenomes	72,795
# base pairs	22.4 Tbp
# of sequences	206.07 billion
# of public metagenomes	12,243

Meyer et al., BMC Bioinformatics, 2008

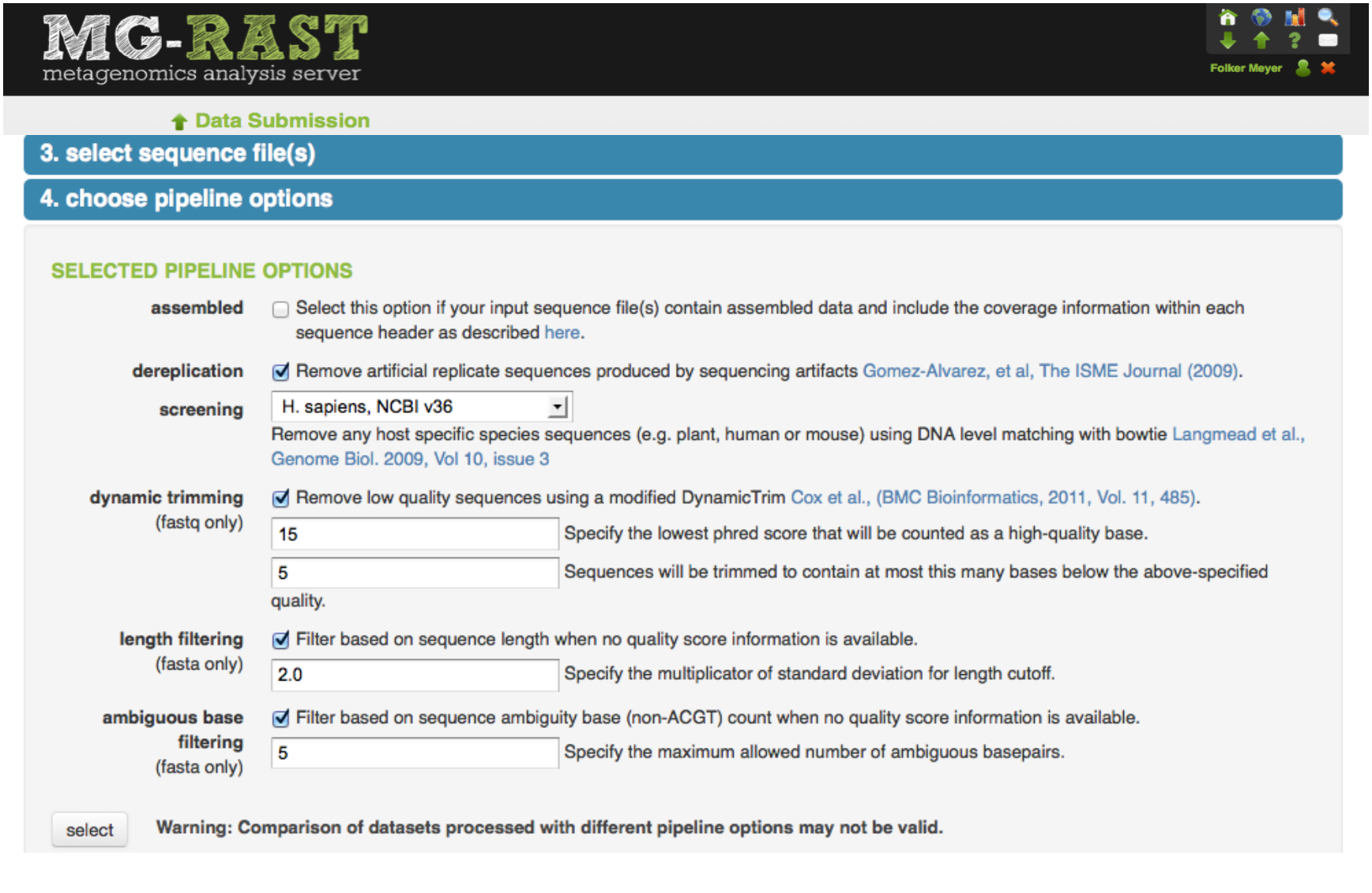
What problems do we solve?

- **Analysis** of single shotgun metagenomics
- With “large data”
  - Don’t tell any physicist I said that
- **Comparison** of MANY data sets
  - **Note:** analyzed data is 10x the size of input data

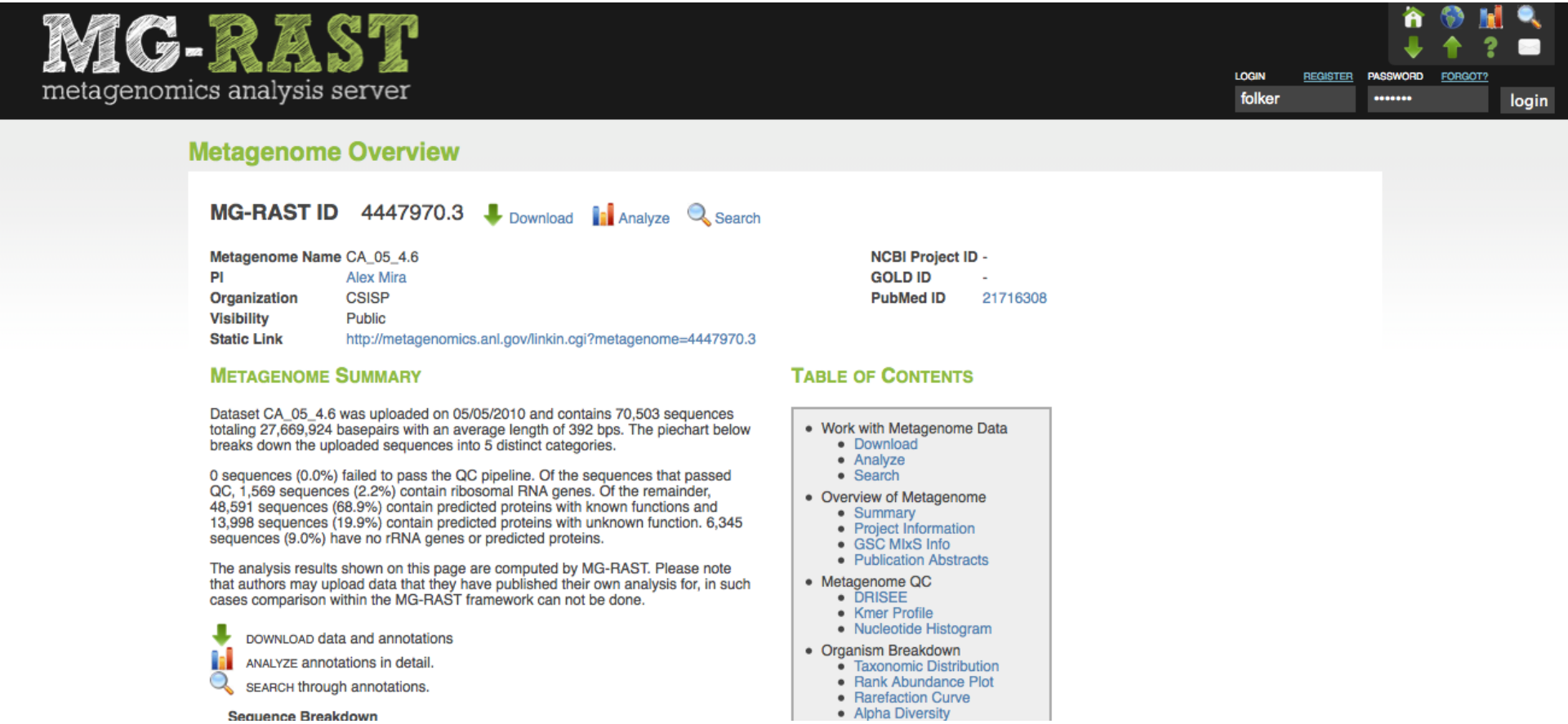
- **We need to be very efficient with resources**
  - We do not use big iron!
  - Had to improve pipeline 750x over past 18months

➔ Triggered infrastructure research projects (SHOCK and AWE)

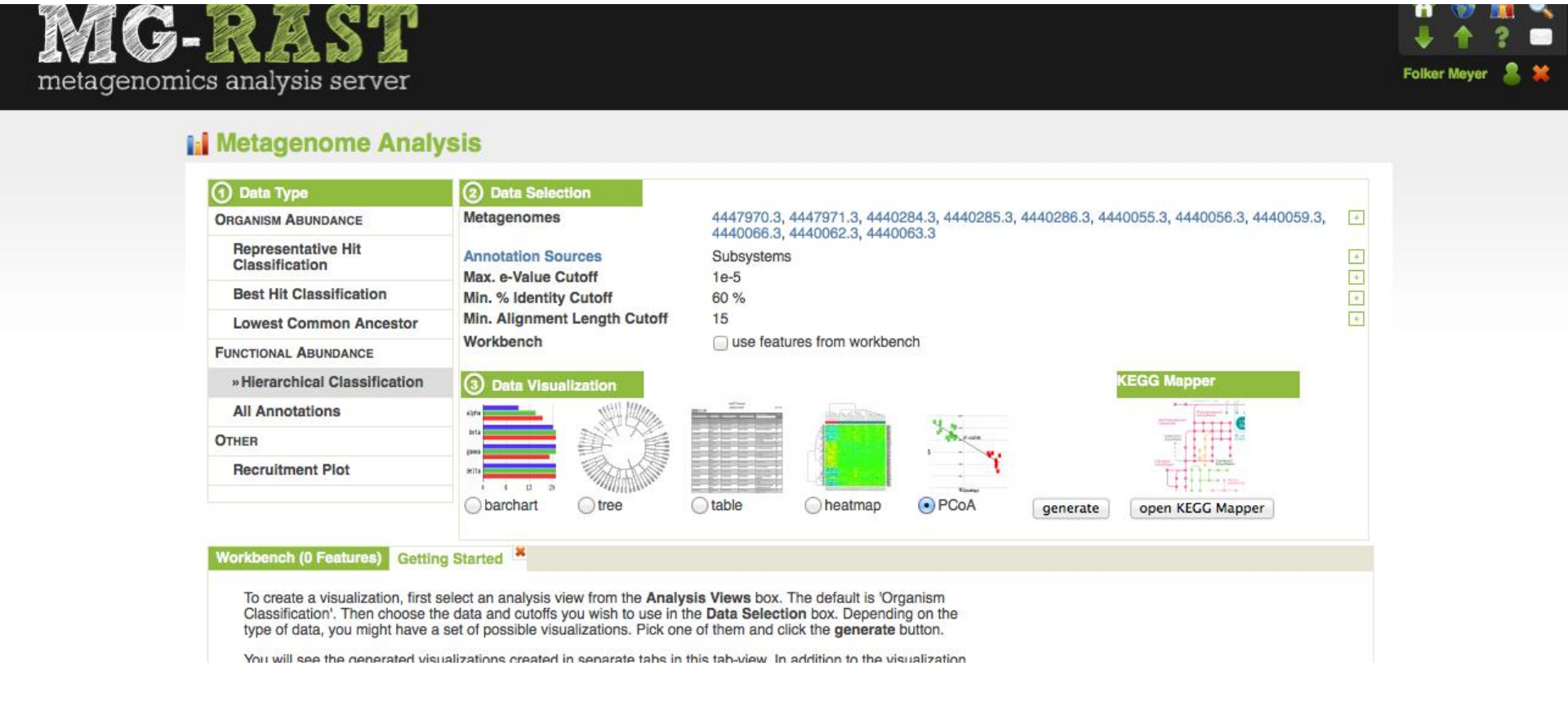
Metagenome Upload (ftp, http, gridFTP)



Metagenome Overview



Comparison tools



MG-RAST
metagenomics analysis server

Folker Meyer

Workbench (0 Features) | Getting Started | Functional PCoA 1 | Functional PCoA 2

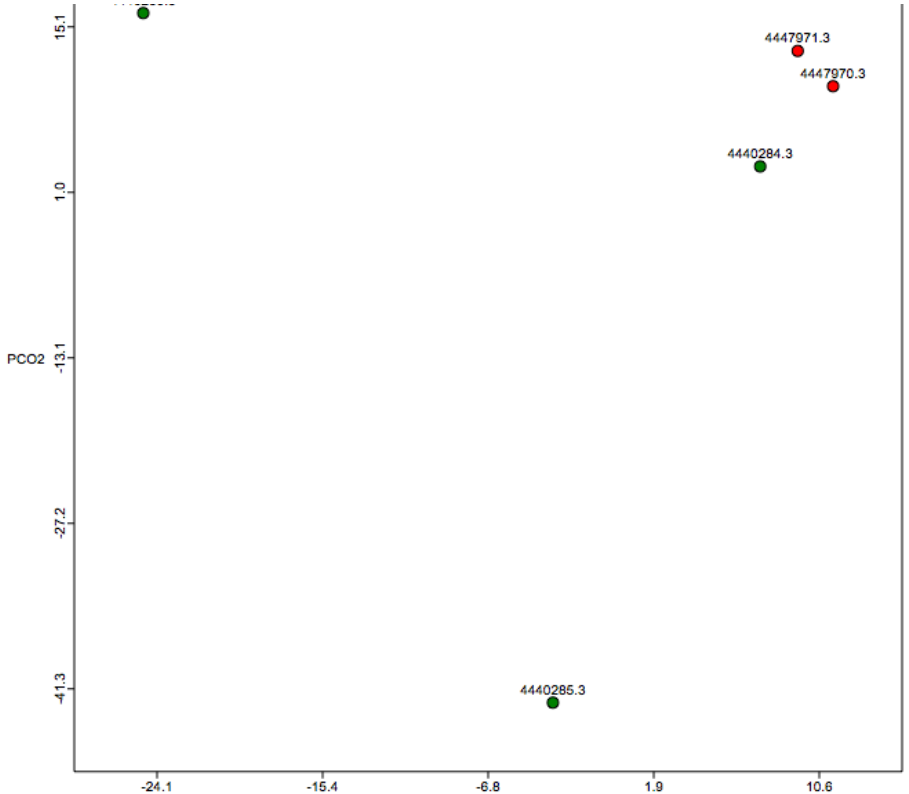
This data was calculated for metagenomes 4447970.3, 4447971.3, 4440284.3, 4440285.3 and 4440286.3. The data was compared to Subsystems using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15 measured in aa for protein and bp for RNA databases.
The data has been normalized to values between 0 and 1. If you would like to view raw values, redraw using the form below.

redraw using normalized values and bray-curtis distance draw

The image is currently dynamic. To be able to right-click/save the image, please click the static button static

PCoA grouping control

download values used to generate this figure



component		x-axis	y-axis
PCO1	0.61620	●	○
PCO2	0.23509	○	●
PCO3	0.13954	○	○
PCO4	0.00918	○	○
PCO5	0.00000	○	○

group	name	save as collection
group 1	human	save
group 2	chicken	save
group 3		save
group 4		save
group 5		save
group 6		save
group 7		save
group 8		save
group 9		save
group 10		save

MG-RAST Notebooks / **IPython** integration

MG-RAST metagenomics analysis server

MGNB - The MG-RAST Notebook

Notebook 1 ✖ +

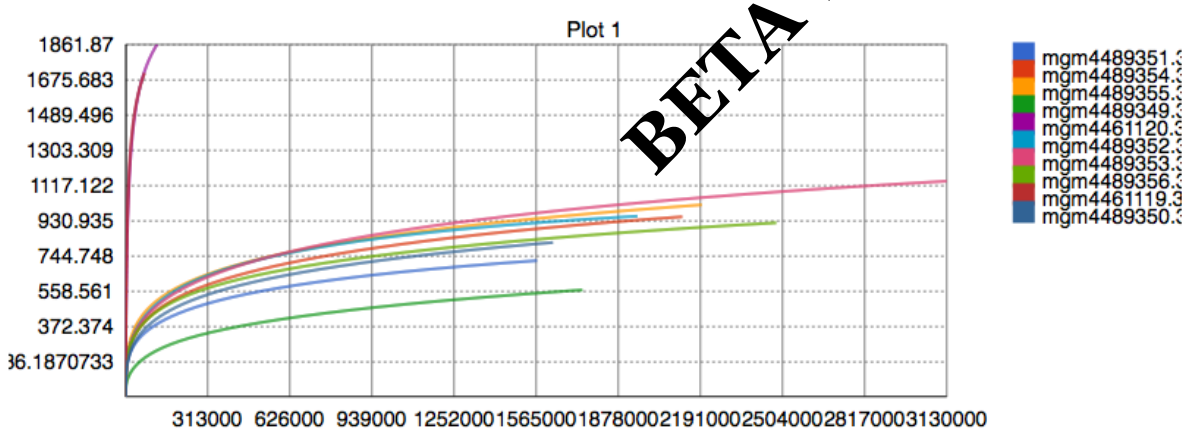
Analysis Builder

File Edit Insert Cell Kernel Help

In [17]:
```
# my data subselection
selected_ids = [ 'mgm4461119.3', 'mgm4461120.3', 'mgm4489349.3', 'mgm4489350.3', 'mgm4489351.3', 'mgm4489352.3', 'mgm4489353.3', 'mg
rare_args_1 = sample_set['statistics'].plot_rarefaction(mgids=selected_ids, arg_list=True)
```

In [18]:
```
rare_args_1.update({'show_legend': True, 'legend_position': 'right', 'connected': True, 'show_dots': False, 'title': 'Plot 1'})
Ipy.RETINA.plot(**rare_args_1)
```

BETA TEST



Data access / Downloads

MG-RAST
metagenomics analysis server

Folker Meyer

Project Overview

THE ORAL METAGENOME IN HEALTH AND DISEASE (ID 128)

metagenomes project metadata

Visibility Public
Static Link http://metagenomics.anl.gov/linkin.cgi?project=128

Share Project | Add Jobs | Edit Project Data | Upload Info | Upload MetaData | Export MetaData

MG-RAST ID	Metagenome Name	bp Count	Sequence Count	Biome	Feature	Material	Location	Country	Coordinates	Sequence Type	Sequence Method	Download
		<	<	huma	human	human-				WGS	454	
4447943.3	CA_04P	142,374,233	339,503	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447192.3	NOCA_01P	77,538,485	204,218	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447103.3	CA1_01P	203,711,161	464,594	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447102.3	NOCA_03P	100,125,112	244,881	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447101.3	CA1_02P	129,851,692	295,072	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447971.3	CA_06_1.6	37,519,874	97,722	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447970.3	CA_05_4.6	27,669,924	70,503	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447903.3	CA_06P	123,266,763	306,740	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis

Download // data access

- Data for all but one visual can be downloaded as spreadsheet
- arbitrary subset can downloaded
  - In standard formats
- All data is available
  - For public projects
- Data is typically private
  - User decides to publish

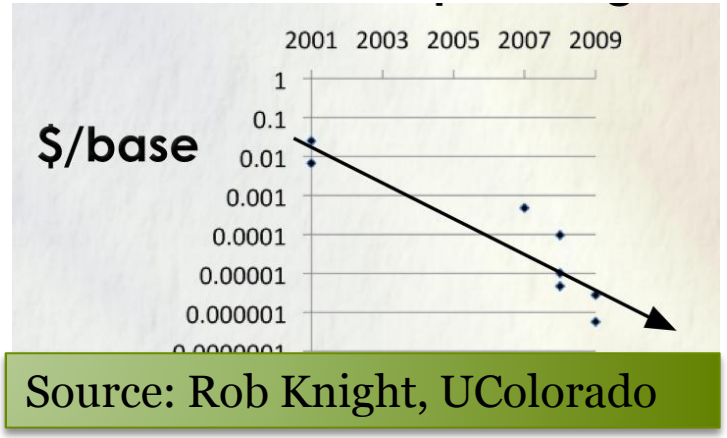
- **Pre-publication** sharing via email-token

Reducing Terabytes to manageable objects

Terabases → 100,000 protein functions with abundance → 4 profiles (COG, SEED, KEGG, NOG)
→ 900,000 organisms → 1 profiles (NCBI taxonomy)
→ ability to go back to sequences

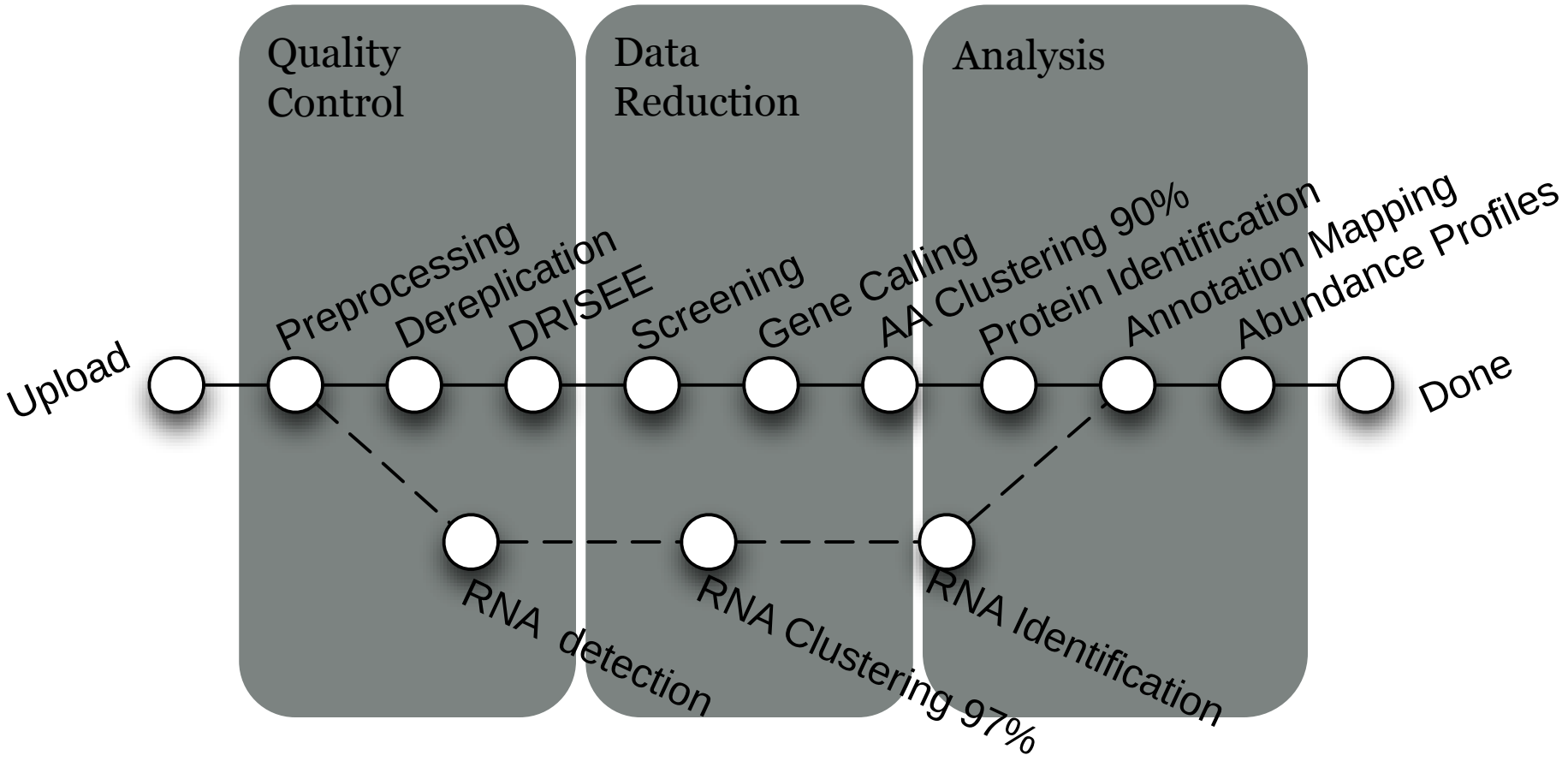
- Data volume growing fast:
  - 2003: C. Venter's GOS with **600MBp** (or 0.6GBp)
  - 2011: HMP with **6TBp** (or 6,000GBp)
- Data is different: shorter reads, but many reads, noisy
- Major cost is in bioinformatics (10x the sequencing cost today)

Driving force

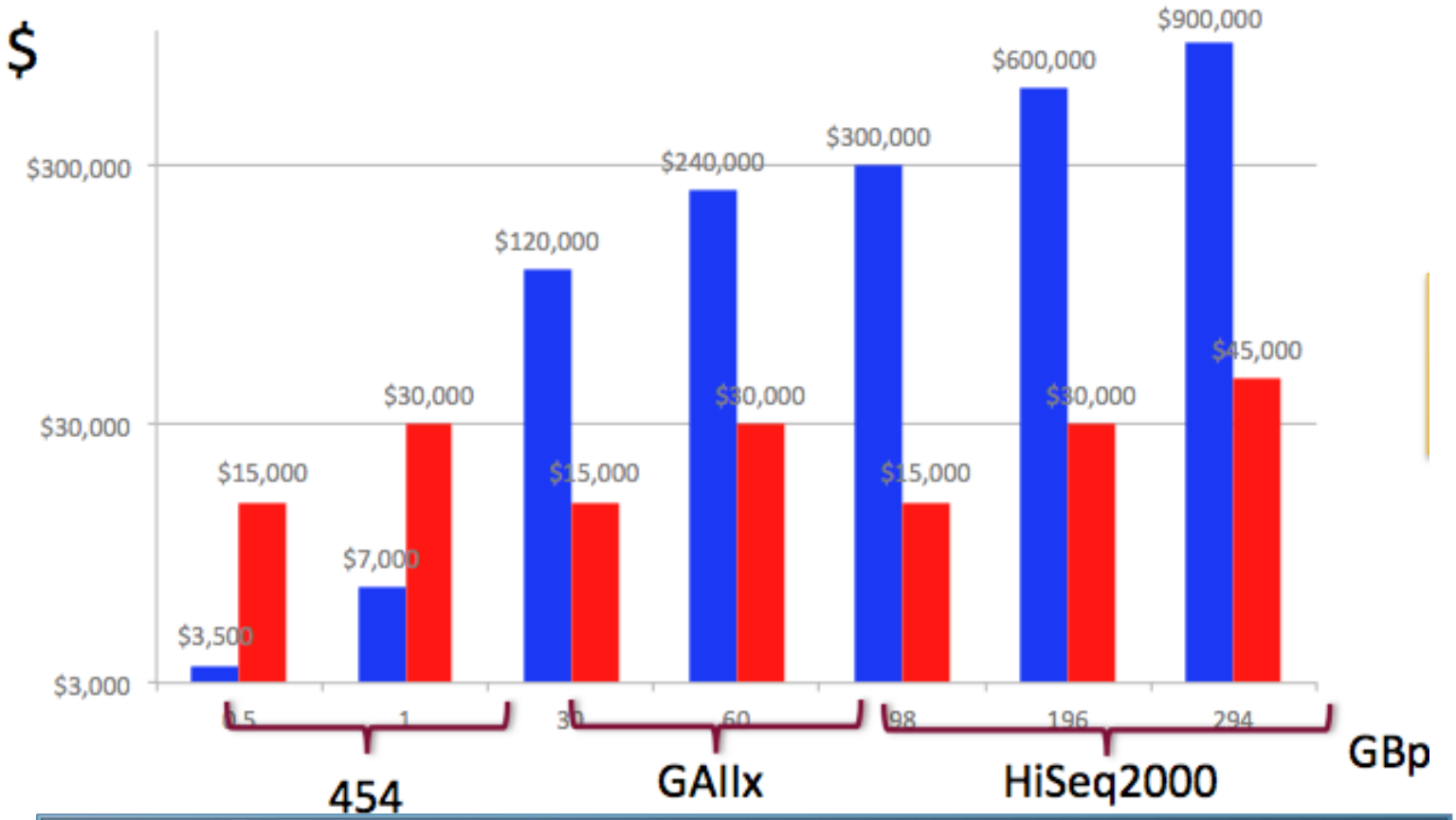


Source: Rob Knight, UColorado

The MG-RAST v3.x pipeline



Analysis cost are dominating



Computing cost (blue) dominate sequencing (red)
- Cost on Amazon EC2 Cloud, September 2009
- Pure run-time for BLASTX, no storage or data transfer

From: Wilkening et al., IEEE Cluster09, 2009

No one size fits all

- Every experiment is different
- Every sample is different

➔MG-RAST does not provide one size fits all
  - No single download for all purposes.

- We allow users to change parameters for analysis at view time
  - Using intelligent data products to construct annotated reads for specific parameters

➔Web interfaces ties it all together

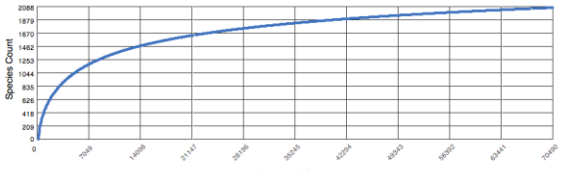
Common views

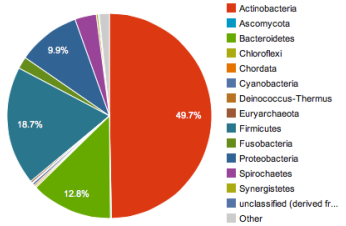
Rarefaction

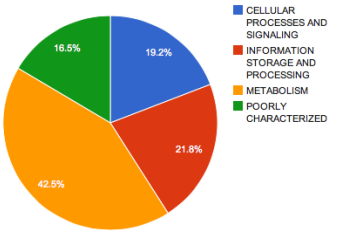
Taxonomic breakdown

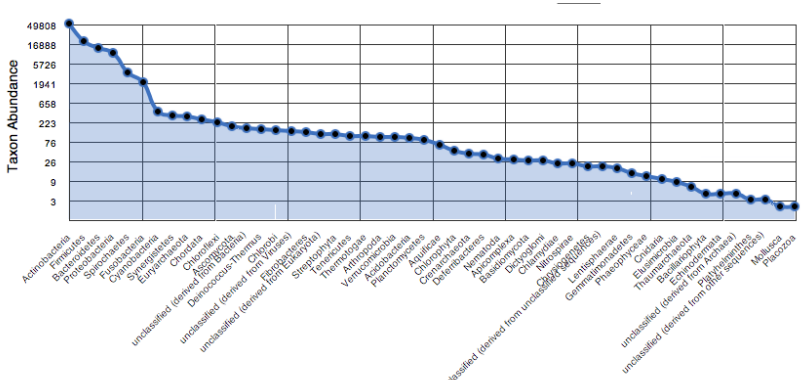
Functional breakdown

Rank abundance



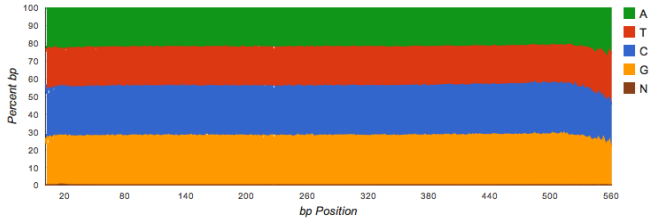


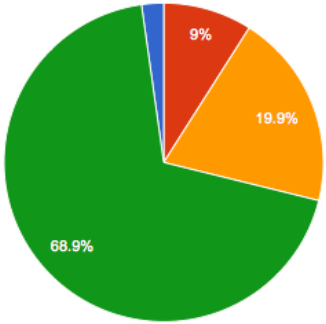


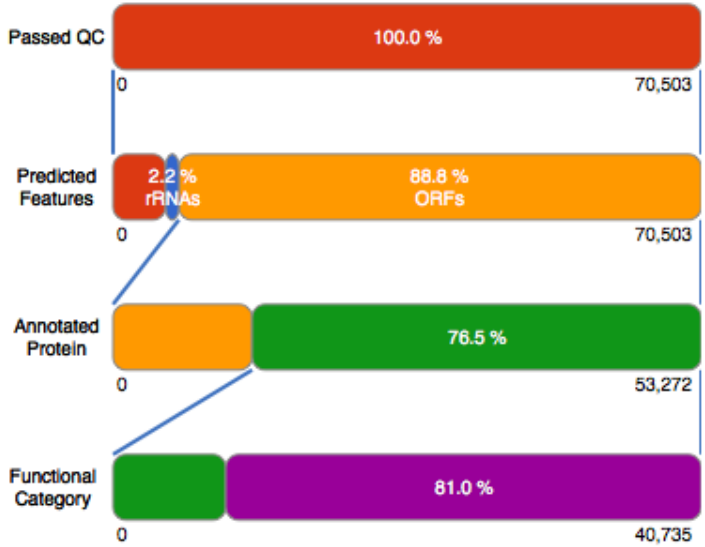


More views

- Nucleotide histogram
- QC
- Processing





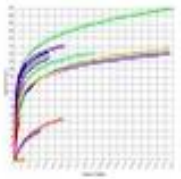


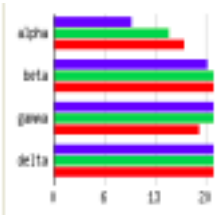
Comparative tools

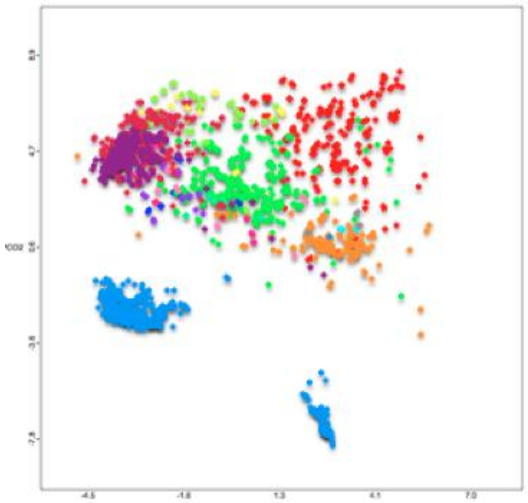
- Comparison of many data sets

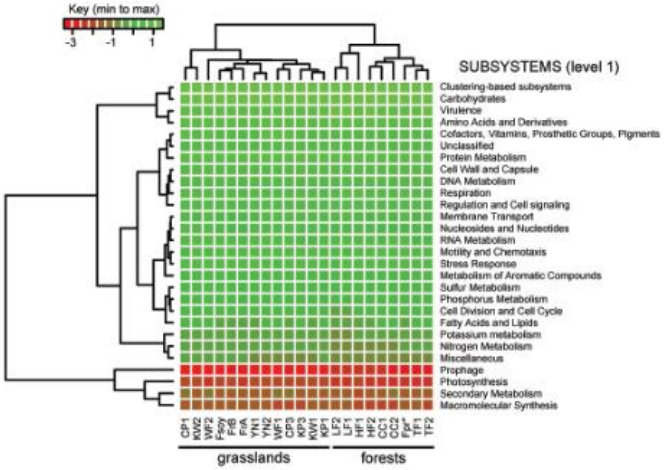
➔**Normalization** allows comparison

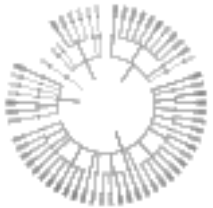
  - Many name spaces (GenBank, SEED, KEGG, GO, COG, eggNOG, UniProt supported)
  - Parameters can be varied at query time











Lessons learned…

- Old style:
  - “Lets sequence as much as we can afford”
  - “Metagenomics is like genomics”

- Today:
  - Often 16s amplicon study first
  - replicates (biological and technical)
    - “design for statistics”
    - “replicate or lie” (Jim Prosser)

  - metadata
    - Genomics Standards Consortium provides tools

  - Provide good QC
    - Identify signal vs. noise ratio
    - Throw away bad data when needed (!)

  - Identify appropriate analysis workflow
  - Perform assembly?

➔ Design for statistics

➔ Metadata (r)evolution

➔ Data hygiene

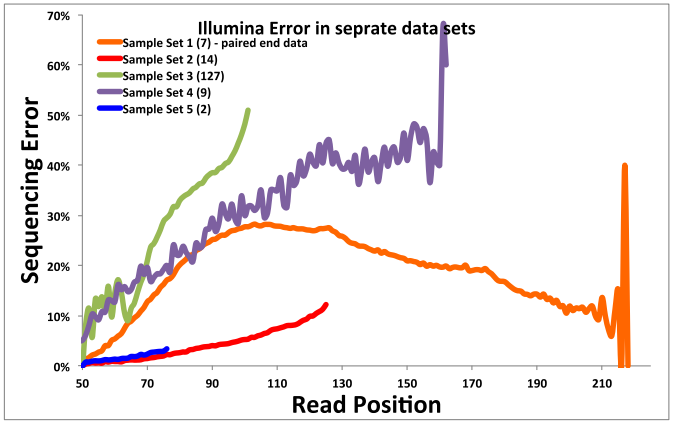
➔ Tool chain matters

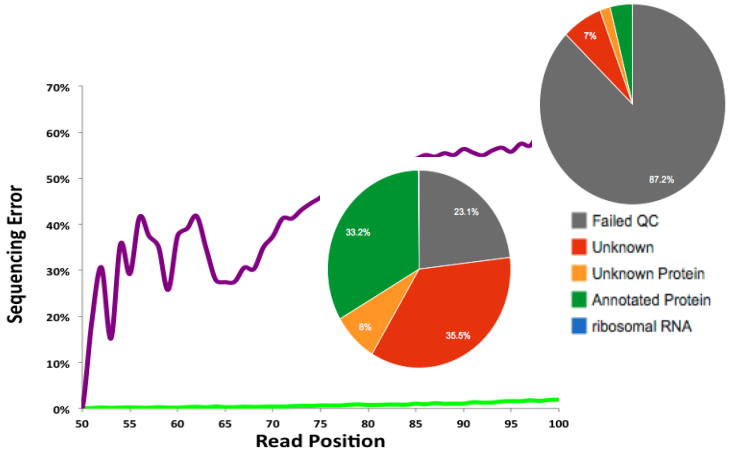
From: Knight, et al, Nature Biotechnology, 2012

DRISEE – objective QC for NGS data

- Approach is simple
  - Develop synthetic reference
- DRISEE:
  - using ADRs to find noise
  - Correlates well with our ability to analyze data

➔Numerous quality issues
  →Not all NGS data is alike
  →Even from the same vendor



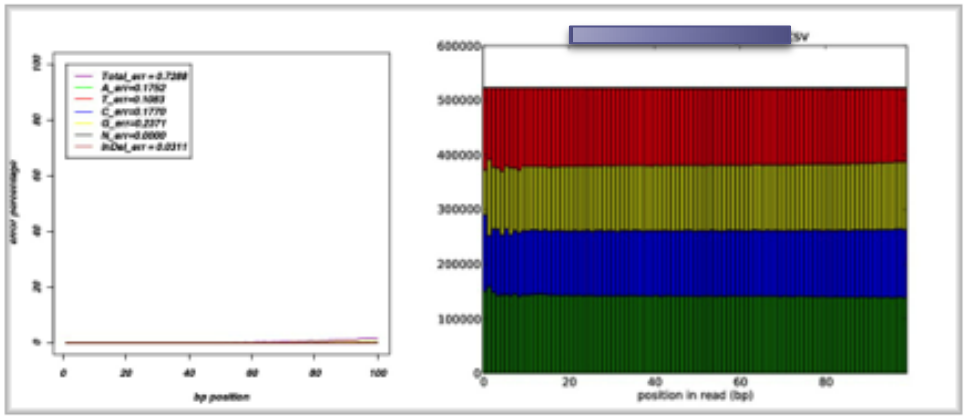


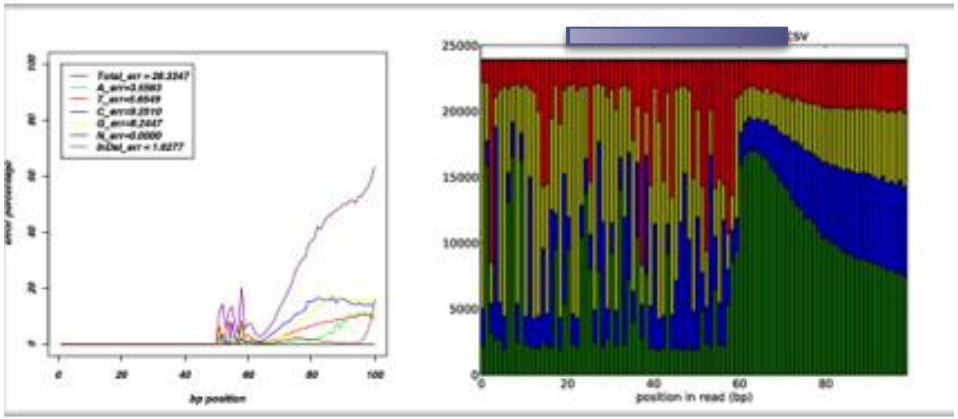
Find prefix identical subsets

➔ **Discrepancies are error**

Keegan et al, PLoS Comp Bio, 2012

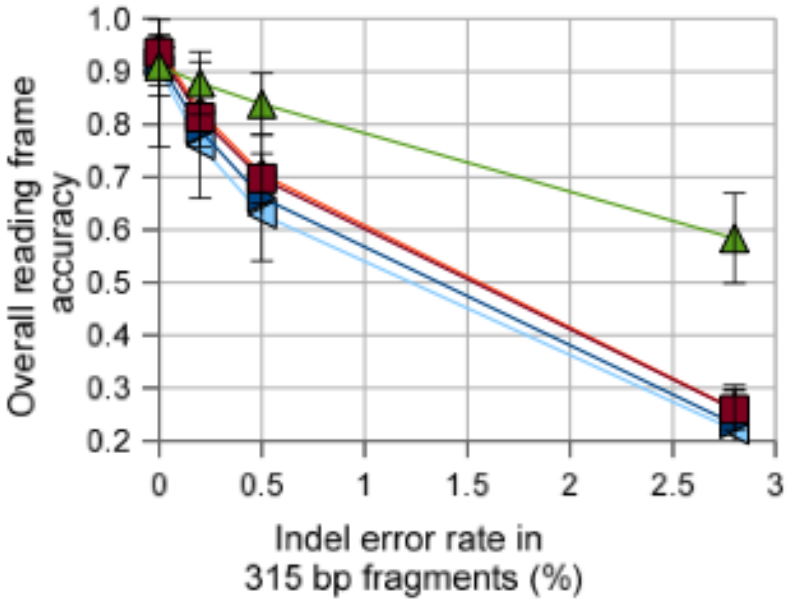
Extreme cases of good and bad data quality





Tool chain variation

- Question:
  - What happens if I vary the tool chain?
- **Existing approaches** rely on:
  - Compare results of different studies (ie multiple pipelines)
- Here we study 5 different popular gene finding tools for metagenome on **simulated data**
- **Effects are dramatic**
  - Accuracy goes drops dramatically with moderate error
- Comparison of data requires identical tool chain



From: Trimble et al, BMC Bioinformatics, 2012

Analysis tool chain

- (small) changes in the analysis tool change will have dramatic impact on results
- Comparing the results of two independently analyzed studies is next to impossible
  - We need to “normalize” analysis
  - MG-RAST currently has 72,000 normalized metagenomes for comparison
  - Over ~11k are public

MG-RAST
metagenomics analysis server

From: Meyer, et al, BMC Bioinformatics, 2008

Programmatic interface (API)

- Part of the KBase API
  - Contributes some unique features
- Uses standards whenever possible
  - Metadata (GSC MIxS)
  - Abundance profiles (BIOM)
- Enables extension by third parties
- Data download and subselection

API Details

- Added 1600 function calls
- Grouped into 14 higher level objects
- Currently access to 11,000 public metagenomes
- Subsetting by function and taxonomy
- Supports KEGG, SEED, UniProt, GenBank, IMG, COG, eggNOG, RDP, SILVA similarities

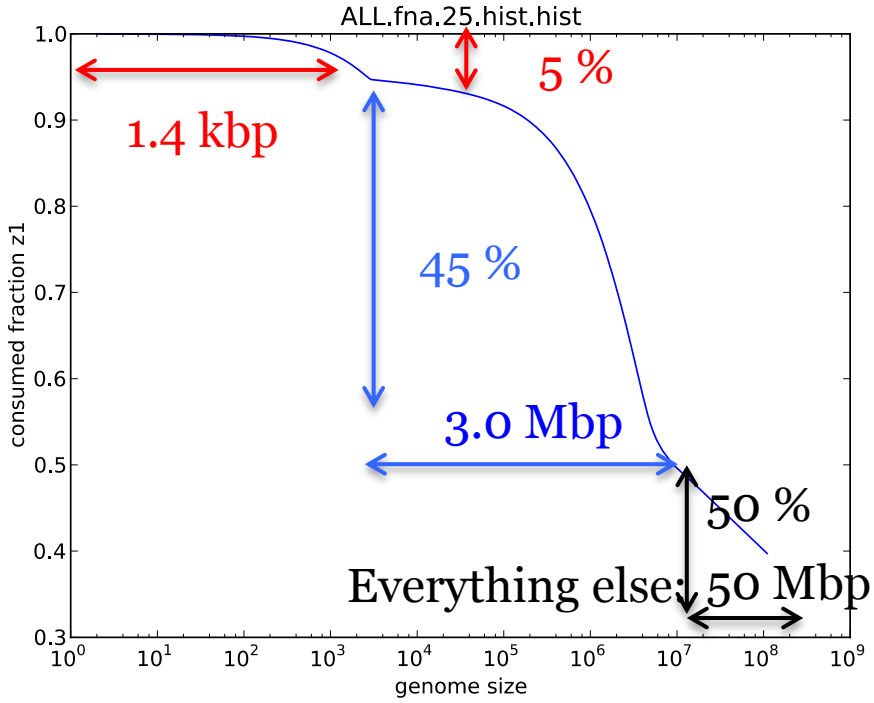
Download (ftp) and RESTful

Status: Public beta

Predicting replicons before assembly

- Using **K-mer spectra** to predict (pan-) genome size
  - K-mer= unique word, easily computed
- In addition to alpha diversity
  - 300 OTU data set
- Using k-mer size 25

- **Red** and **blue** replicons were missing in assembly
  - Allows adjustment of parameters

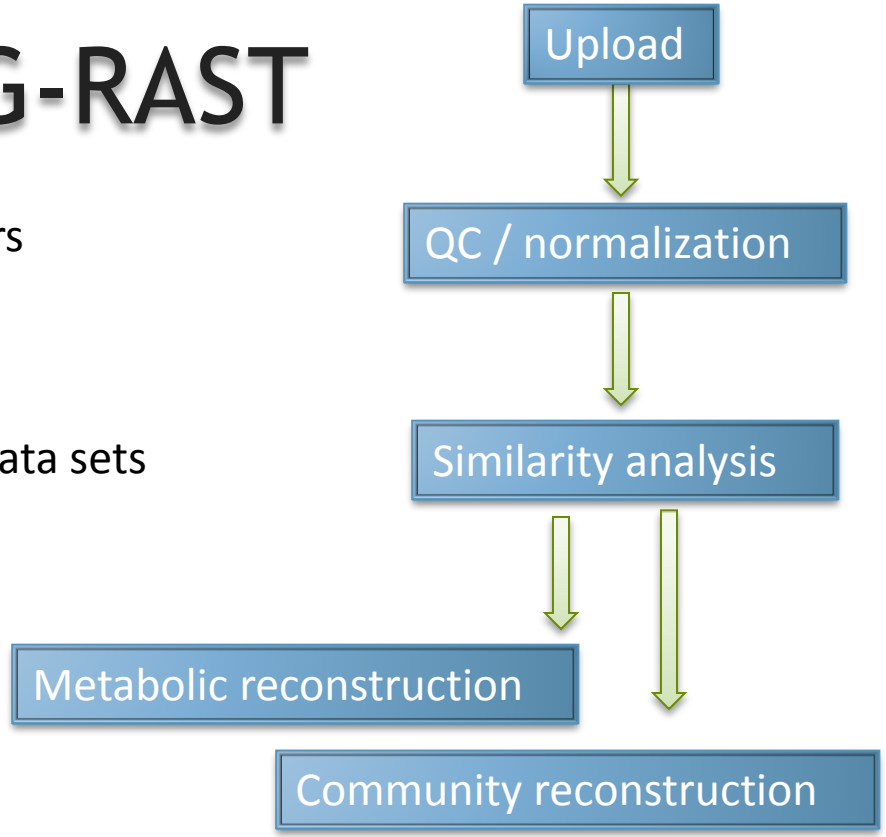


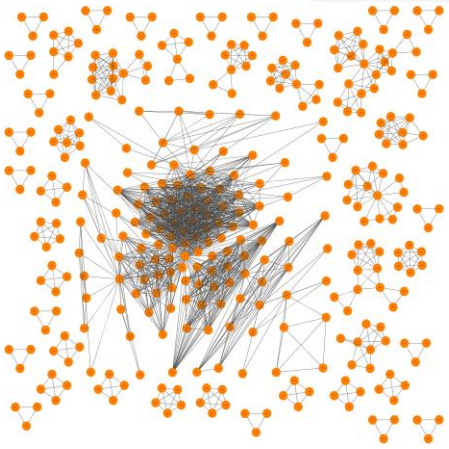
From: Trimble, et al, in prep

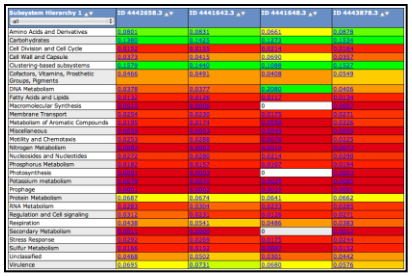
Thank you very much for you attention

Brief history of MG-RAST

- December 2007 (v1)
  - 100+ groups and ~250 data submitters
  - 100+ data sets, ~10+ GBp total size
- October 2009 (v2)
  - **Pre-publication sharing available**
  - ~1500 data submitters, ~300 public data sets
  - 6000+ data sets
  - 200+ GBp total data set size
  - About ~30 GBp/month throughput
- March 2011 (release v3)
  - 2500+ data submitters
  - 25,000 data sets total
  - 3000 submissions in 24h
- February 2012 (v 3.1.2)
  - 10 Terabasepairs (10^12 bp)
- March 2013 (v3.3)
  - 22 Terabasepairs
  - 72k data sets







simplified

Meyer et al., BMC Bioinformatics, 2008